Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs

Habib Benali

INRIA

CENTRE DE RENNES

IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France

Tél.(1) 39 63 55 11

Rapports de Recherche

N° 528

STABILITÉ DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES EN CAS DE DONNÉES MANQUANTES ET DE MODALITÉS A FAIBLES EFFECTIFS

Habib BENALI

Mai 1986

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042-RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Publication Interne n° 294 - Avril 1986 - 16 pages

Habib BENALI - IRISA RENNES

STABILITE DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES EN CAS DE DONNEES MANQUANTES ET DE MODALITES A FAIBLES EFFECTIFS

Introduction

Une des sources de perturbation des résultats d'une analyse des correspondances mutiples (A.F.C.M.) est le problème des non-réponses et des réponses rares. Une variante de cette méthode proposée par B. ESCOFIER [3] permet de résoudre ce problème. Un programme de cette technique présentant la même structure que MULTM de SPAD [5] a été écrit. Certains sous-programmes d'analyse des correspondances multiples de MULTM ont été repris et adaptés sans changer profondément l'algorithmique. Nous comparons dans cet article les résultats de cette variante à la technique utilisée dans SPAD où les réponses rares sont ventilées au hasard sur une des autres modalités de réponses à la même question et où les réponses manquantes sont considérées comme réponses particulières.

Abstract

One source of perturbation in the results of multiple correspondance analysis, is the problem of the none-responses and rare responses. A variant of multiple correspondance analysis proposed by B. ESCOFIER [3] allows as to resolve this problem. A programme about this technique presenting the same structure than SPAD'S MULTM was written. Some MULTM multiple correspondance analysis susprograms were revised and adapted. On this paper, we compare the results of this variant to the technique used in SPAD where rare responses are randomly distributed on another same question response modality, and where missing responses are considered as particular ones.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 I. N. S. A. DE RENNES

INSTITUT NATIC
EN INFORMATIQU
(LABORATC

NOTATIONS

Un questionnaire est formé d'un ensemble Q de questions dont chacune admet un ensemble J_q de modalités de réponses. On note :

I l'ensemble des individus

J l'ensemble des modalités de réponses à toutes les questions Q

K_{IJ} le tableau disjonctif des variables indicatrices associées aux modalités de réponses

$$K_{IJ} = [K_{IJ}, K_{IJ_2}, \ldots, K_{IJ_Q}]$$

card I = n

card J = card J_1 + card J_2 +...+ card J_Q

$\forall j \in J_q \subset J$ $\quad K_{ij} = 1$ si l'individu i possède la modalité j

ou

$K_{ij} = 0$ sinon

$$K_{i.} = \sum_{j \in J} K_{ij}$$

$$K_{.j} = \sum_{i \in I} K_{ij}$$

$$K = \sum_{i,j} K_{ij}$$

En A.F.C.M., le tableau K_{IJ} est disjonctif complet, $K_{i.} = Q = \frac{K}{n}$. Les programmes de calculs utilisent le tableau du codage condensé C_{IQ} du tableau K_{IJ} où la case (i,q) contient le numéro c_{iq} de la modalité de la question q choisie par l'individu i. La mise sous forme disjonctive des données dans cette article n'est qu'une présentation commode.

Un individu i est représenté dans R_J par son profil ligne $\{\frac{K_{ij}}{K_{i.}}, j \in J\}$, et une modalité j est représentée par son profil colonne $\{\frac{K_{ij}}{K_{.j}}, i \in I\}$.

Le nuage des individus N(I) est l'ensemble des profils lignes affectés des poids $\frac{K_{i.}}{K}$.

Le nuage des modalités N(J) est l'ensemble des profils colonnes affectés des poids $\frac{K_{.j}}{K}$.

METHODES UTILISEES DANS SPAD

Cette méthode préserve la structure du tableau disjonctif complet, en ventilant au hasard les réponses rares des individus sur une des modalités de réponses à la même question, et considérant comme modalités particulières les réponses manquantes.

Si la réponse manquante correspond à une attitude particulière de non-réponse, ceci est justifié. Nous nous intéressons ici au problème de données manquantes non significatives d'une attitude.

METHODE DU CUMUL

Pour résoudre le problème des modalités rares, cette technique consiste à les cumuler en une colonne reste et analyse le tableau disjonctif des données contenant cette colonne ; ce qui permet de garder la marge constante sur I.

Cette méthode a été testée (1), et ses résultats présentent de fortes variations. Nous n'en dirons plus rien ici.

METHODE MULMD

Cette méthode est une variante de l'A.F.C.M., elle résoud simultanément le problème des réponses manquantes et des réponses rares, en minimisant leur influence sur le résultat, elle traite des tableaux disjonctifs "incomplets" ainsi définis :

Tableau disjonctif incomplet $K'_{IJ'}$

Dans le tableau K_{IJ}, les non-réponses à une question q sont codées zéro sur l'ensemble des modalités J_q de cette question, et les colonnes correspondant aux modalités rares sont supprimées.

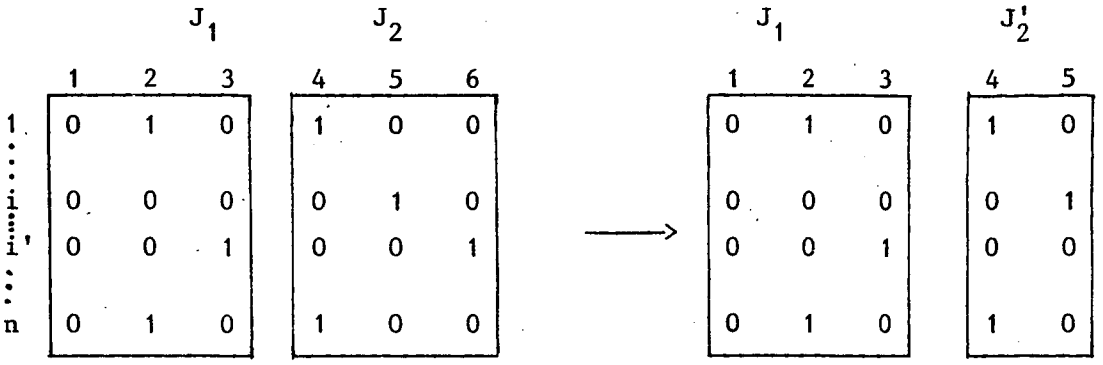

Tableau 1 Tableau 2

i ayant une non-réponse à J_1, elle est code zéro, i' ayant pris la modalité 6 de J_2, qui est rare et qui est supprimée dans le second tableau. On remarque que la marge sur l'ensemble des individus dans le tableau disjonctif incomplet n'est pas constante.

Choix d'une distance pour l'étude du tableau disjonctif incomplet K'_{IJ}

Distance du KHI-DEUX

L'A.F.C.M. classique utilise la distance du KHI-DEUX entre deux profils lignes (resp. colonnes) qui s'écrit :

$$d^2(i,i') = \sum_{j \in J'} \frac{K}{K_{.j}} \left[\frac{K_{ij}}{K_{i.}} - \frac{K'_{ij}}{K_{i'.}} \right]^2$$

$$d^2(j,j') = \sum_{i \in I} \frac{K}{K_{i.}} \left[\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j'}} \right]^2$$

Inconvénient de cette distance

Dans un tableau disjonctif incomplet, le problème des modalités rares (contribuant fortement à la distance entre deux profils lignes) est résolu. Par contre, si deux individus i et i' n'ont pas donné le même nombre de réponses ($K_{i.} \neq K_{i'.}$), une modalité j choisie simultanément par ces individus augmente leur distance car le terme $\frac{K_{ij}}{K_{i.}} - \frac{K_{i'j}}{K_{i'.}}$ n'est pas nul, ce qui est un inconvénient et pose un réel problème d'interprétation. Cette métrique est donc inadaptée à l'étude de tableau disjonctif incomplet.

Distance variante du KHI-DEUX

Pour remédier à cet inconvénient, on remplace la marge $(K_{i.}, i \in I)$ par la marge constante $(\frac{K}{n}, i \in I)$ partout où elle intervient : profil et poids des individus, métrique et origine des axes pour le nuage des profils colonne.

Les distances entre profils lignes sont analogues à celles issues du tableau disjonctif complet K_{IJ} obtenu en supprimant les termes provenant des non-réponses et des modalités rares, et les distances entre profils colonnes sont identiques à celles issues du tableau K_{IJ}.

Dualité et formules de transitions

Les facteurs F_s du nuage N(I) se déduisent des facteurs G_s du nuage N(J) par les formules de transitions suivantes où λ_s est la valeur propre d'ordre s.

$$F_s(i) = \frac{n}{\sqrt{\lambda_s} \sum_{j \in J'} K_{.j}} \sum_{j \in J} K_{ij} G_s(j) - \frac{1}{\sqrt{\lambda_s} \sum_{j \in J'} K_{.j}} \sum_{j \in J} K_{.j} G_s(j)$$

$$G(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{K_{ij}}{K_{.j}} F_s(i)$$

Dans la première formule apparaît le terme $\frac{1}{\sqrt{\lambda_s} \sum_{j \in J} K_{.j}} \sum_{j \in J} K_{.j} G_s(j)$ qui représente la coordonnée du centre de gravité G du nuage N(J) sur l'axe F_s, et mesure le décalage du facteur quand l'origine des axes ne correspond pas à G.

Ce terme, nous le verrons en pratique est presque nul, ce qui permet d'interpréter comme en A.F.C.M. classique l'abscisse d'un individu comme le barycentre des modalités de réponses qu'il a choisies. La deuxième formule est exactement celle de l'A.F.C.M.

Elément supplémentaire et formule de reconstitution des données

On peut mettre des modalités en élément supplémentaire, particulièrement les modalités non-réponses et les modalités rares. L'abscisse $G_s(j+)$ d'une modalité supplémentaire est définie par :

$$G_s(j+) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{K_{ij}^{+}}{K_{.j}^{+}} F_s(i)$$

A partir des formules de transition, on peut reconstituer le tableau initial en utilisant la formule :

$$K_{ij} = \frac{K_{.j}}{n} \left(1 + \sum_{s}^{r} \frac{1}{\sqrt{\lambda_s}} F_s(i)\, G_s(j)\right)$$

Nuage N(I) des profils des lignes

La distance entre deux profils lignes i et i' est définie par :

$$d^2(i,i') = \sum_{j \in J} \frac{K}{K_{.j}} \left[\frac{n}{K} K_{ij} - \frac{n}{K} K_{i'j}\right]^2 = \frac{n^2}{K} \sum_{j \in J} \frac{1}{K_{.j}} \left[K_{ij} - K_{i'j}\right]^2$$

L'inconvénient rencontré dans la distance du hhi-deux disparaît. L'analyse du nuage N(I) est faite à partir de son centre de gravité qui sera pris comme origine des axes.

Nuage N(J) des profils des colonnes

La distance entre deux profils colonnes j et j' est définie par :

$$d^2(j,j') = \sum_{i \in I} K_{..} \frac{n}{k} \left[\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j'}}\right]^2 = n \sum_{i \in I} \left[\frac{K_{ij}}{K_{.j}} - \frac{K_{ij'}}{K_{.j'}}\right]^2$$

Cette métrique possède la propriété intéressante qu'est "l'équivalence distributionnelle".

L'analyse du nuage N(J') est faite à partir du point de coordonnée K/n pris comme origine des axes. Les facteurs sur J' ne seront pas exactement centrés puisque le centre de gravité de N(J') est $(K_{i.}/K,\ i \in I)$.

En gardant les modalités rares et en ajoutant des modalités de non réponses pour les réponses manquantes, on obtient un tableau disjonctif complet K''_{IJ} avec $J'' \supset J \supset J'$ de marge sur I constante. Le nuage N(J) des profils des colonnes de K_{IJ}

est un sous nuage de N(J") associé à $K"_{IJ"}$: les profils sont identiques et la métrique de R_I est définie par la marge constante. On analyse alors le sous nuage N(J) du nuage complété N(J") en gardant la métrique et le centre de gravité associé à N(J").

Mise en oeuvre pratique de la méthode

Le programme

Il s'agit du programme MULTM de SPAD (5) adapté aux traitements des questionnaires avec non-réponses et modalités à faibles effectifs. Dans le programme principal, le tableau de Burt d'ordre (J',J') est calculé à partir du tableau du codage condensé C_{IQ}, la matrice sur les profils colonne d'ordre (J',J') est calculée, sa dimension ne peut être réduite (1), ce qui est la grande particularité par rapport au programme original MUTM de SPAD.

Cas des réponses rares

On fixe au seuil NMIN à partir duquel les modalités d'effectifs inférieurs à NMIN sont considérés comme rares, elles sont affectées à un fichier IQUES où elles sont éliminées des calculs.

Cas des réponses manquantes

Les non-réponses sont générées de façon aléatoire ((4),(6)) à partir d'un tableau complet, elles sont codées sur le fichier initial comme les autres réponses ; le nombre de modalités 'réponses manquantes' ainsi créé est fixé par un paramètre NABAND, elles sont lues puis éliminées des calculs.

Comparaison des deux méthodes

L'étude comparative des résultats de la méthode proposée avec celle de SPAD est faite sur deux fichiers différents. Sur le premier de dimension (188,120), on étudie dans un premier temps l'influence des modalités rares en augmentant l'effectif minimum NMIN ; dans un second temps, après avoir fixé un seuil NMIN pour ce type de perturbation, on étudie l'influence des réponses manquantes générées de façon aléatoire (cf. (4),(6)) sur les variables les plus discriminantes du tableau initial en augmentant leur nombre.

Sur le second fichier de dimension (340,52), où il n'y a pas de modalités d'effectifs faibles, on étudie le problème des données manquantes lorsqu'elles sont réparties sur toutes les variables.

Remarque : Les deux méthodes donnent des résultats identiques pour un tableau complet n'ayant aucune modalité d'effectif inférieure à NMIN.

Cas des réponses rares

Une qualité importante d'une méthode supprimant les modalités rares est la robustesse des résultats et la stabilité des configurations par rapport à de petites variations de NMIN.

Tableaux de corrélation entre facteurs de même rang pour deux valeurs successives de NMIN

* NMIN prend les valeurs 2% et 3% de l'effectif total du tableau

Méthode MULMD

	corrélation
Axe 1	0,99
Axe 2	-0,89
Axe 3	0,84
Axe 4	-0,92
Axe 5	0,94

Tableau 3

Méthode de SPAD

	corrélation
Axe 1	0,97
Axe 2	-0,80
Axe 3	-0,70
Axe 4	0,83
Axe 5	-0,88

Tableau 4

Entre deux valeurs successives de NMIN, les facteurs de la variante de l'A.F.C.M. restent plus corrélés. Cette remarque reste vraie dans le cas général où le nombre de modalités rares supprimées augmente. Au vu de ces résultats, la méthode qu'utilise SPAD semble plus sensible que la variante de l'A.F.C.M.

Tableau de corrélation entre facteurs de même rang des deux méthodes pour le même seuil NMIN

* NMIN prend les valeurs 2%, 3%, et 16% de l'effectif total du tableau

	NMIN égal		
	2% de l'effectif total	3% de l'effectif total	16% de l'effectif total
Facteur 1	-0,99	-0,99	-0,96
Facteur 2	-0,97	-0,94	-0,96
Facteur 3	-0,95	0,92	-0,97
Facteur 4	-0,93	0,95	0,73
Facteur 5	-0,86	0,94	-0,11

Tableau 5

Les fortes corrélations des trois premiers facteurs au seuil NMIN égal à 16 % prouvent que les facteurs des deux méthodes varient dans le même sens ; ce résultat intéressant montre que la perturbation induite par les modalités rares a même effet sur ces facteurs. Ceux-ci sont par ailleurs identiques au seuil NMIN égal à 2%, considéré comme seuil de stabilité.

Cas de réponses manquantes

Le problème des réponses manquantes est plus délicat lorsque les variables les plus discriminantes d'une analyse se trouvent touchées. On étudie la stabilité pour ce type de perturbation sur le premier fichier (188,120) après avoir fixé le paramètre NMIN à 2% de l'effectif total de l'échantillon. On génère 5% de données manquantes sur une, trois puis dix variables les plus discriminantes et enfin on augmente ce seuil à 20% pour les dix variables seulement. On retient les cinq premiers facteurs de chacune des analyses dans chaque méthode et on effectue une Analyse en Composantes Principales sur l'ensemble (4x5=20) de ces facteurs. L'inertie expliquée par les cinq premiers facteurs de l'A.C.P. est de 95,17 %, ce qui laisse entrevoir une "certaine" perturbation des résultats des différents A.F.C.M.. Pour visualiser cette perturbation, on représente sur les premiers plans factoriels les différents facteurs.

Notations sur les graphiques

M désigne les facteurs de la méthode MULMD et S ceux de la méthode utilisée par SPAD.

On note MI (resp. SI) les facteurs issus des tableaux sans réponses manquantes, et MIJ (resp. SIJ) les facteurs des tableaux perturbés ; où I désigne le rang du facteur (I=1,5), et J la perturbation sur le tableau initial (J=1,4).

J=1 : 5% de données manquantes sur une variable
J=2 : 5% de données manquantes sur trois variables
J=3 : 5% de données manquantes sur dix variables
J=4 : 20% de données manquantes sur dix variables.

Ainsi, M2 correspond au 2ème facteur (sans réponses manquantes) obtenu par MULMD.

S13 correspond au 1er facteur obtenu dans SPAD avec 5% de données manquantes sur 10 variables.

On note sur les différentes figures que le décalage entre les représentations des facteurs de même rang est moins important dans la variante de l'A.F.C.M. où tous les points se trouvent regroupés autour du point stable ; une variation du facteur 3 apparaît (FIG3) quand le seuil de non réponses atteint 20% dans les dix variables les plus discriminantes.

Cette méthode assure une robustesse aux résultats, car il est assez rare de se trouver dans le cas où 20% des données manquent dans les dix variables les plus importantes.

Au vu de ces premiers résultats, on a généré sur le fichier (340,52) ne comportant aucune modalité rare (où l'analyse par la technique de SPAD et celle de la variante ont donné les mêmes résultats) 10% de non-réponses réparties de façon aléatoire sur l'ensemble des cases du tableau de données.

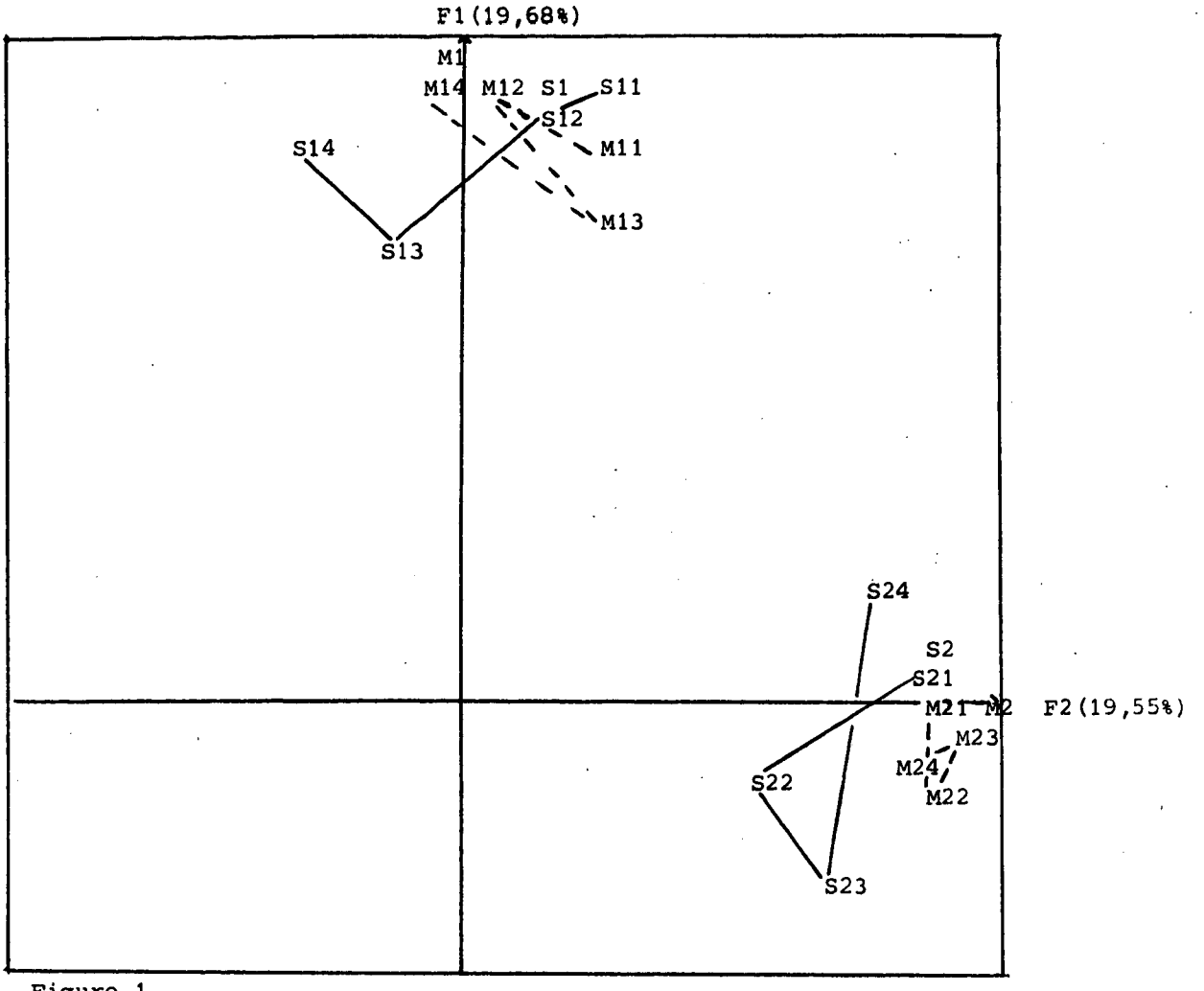

Figure 1

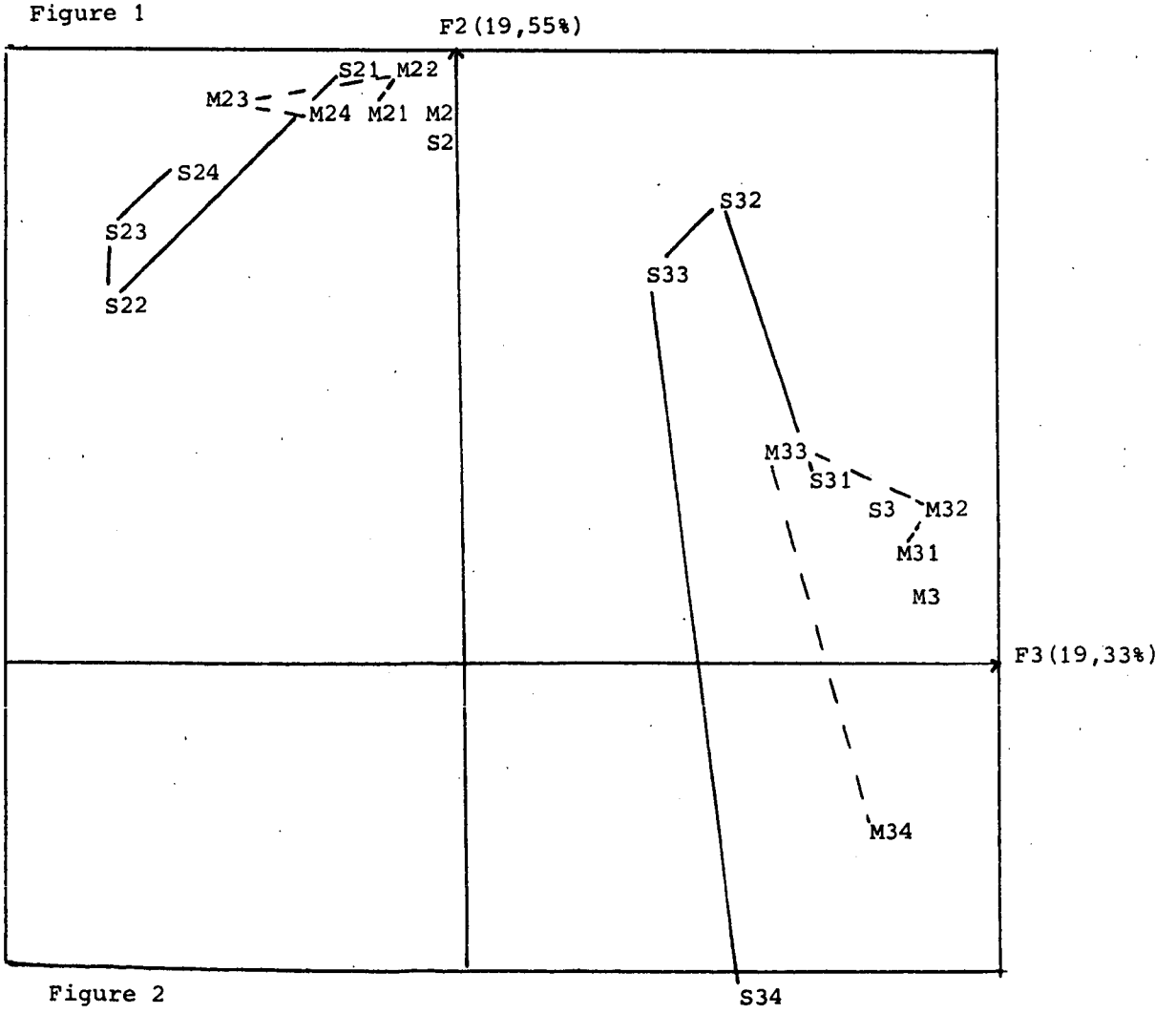

Figure 2

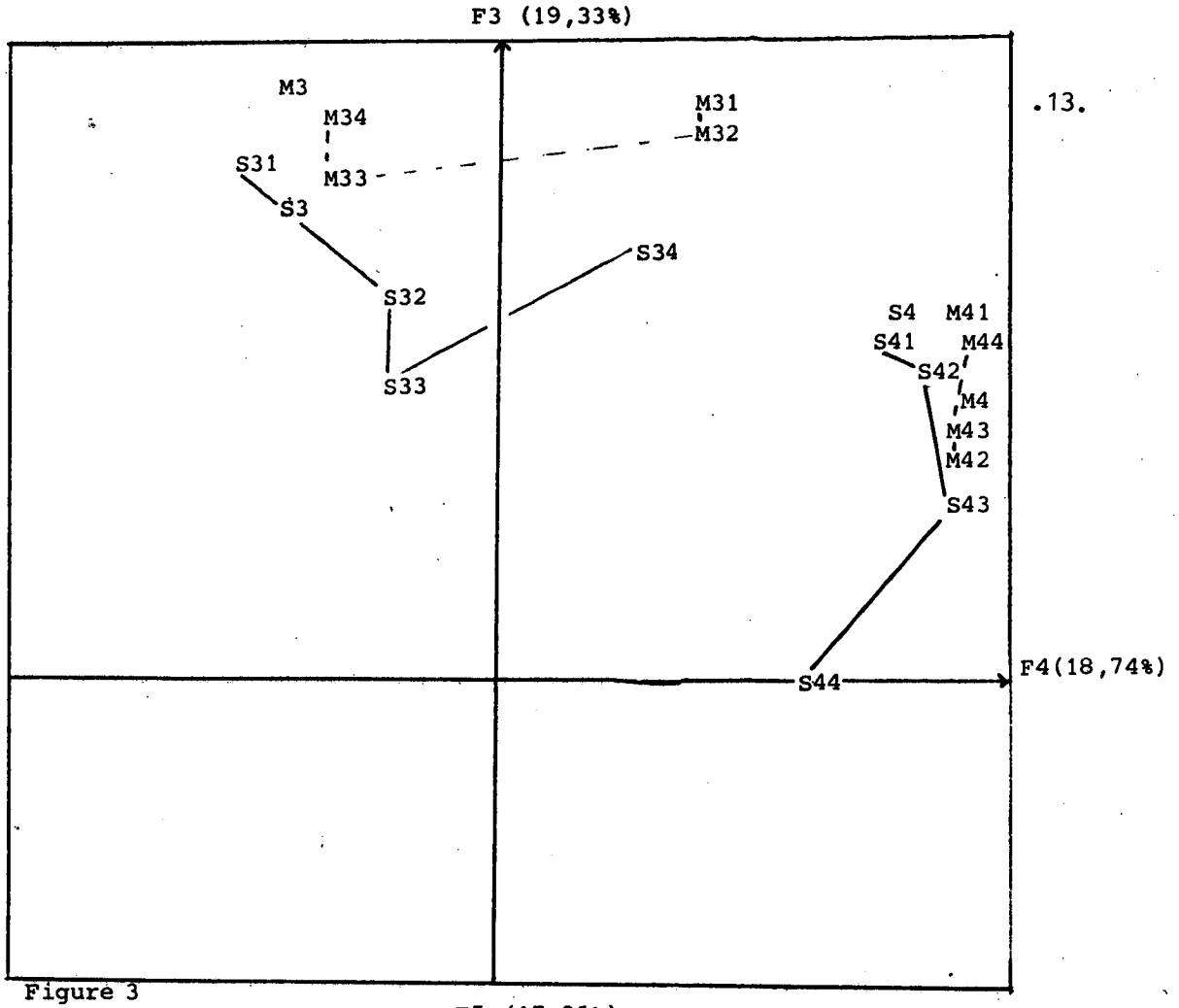

Figure 3

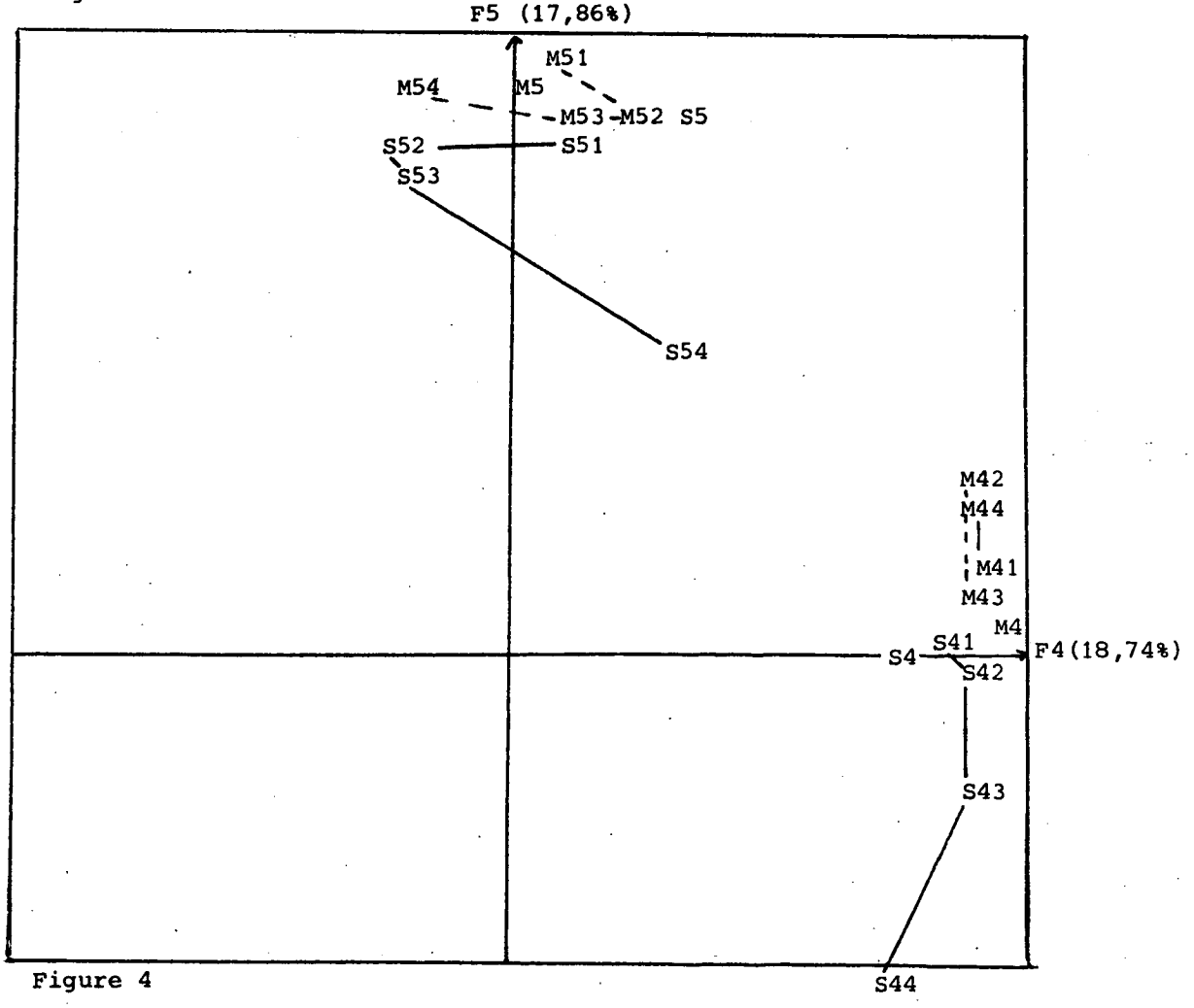

Figure 4

Comparaison des facteurs de même rang des deux méthodes

Les tableaux suivants donnent les corrélations entre facteurs de même rang obtenus respectivement avec et sans non-réponses.

	Analyse par la variante de l'A.F.C.M.
Facteur 1	0,96
Facteur 2	0,98
Facteur 3	-0,95
Facteur 4	-0,95
Facteur 5	-0,65

Tableau 6

	Analyse par 'SPAD'
Facteur 1	-0,98
Facteur 2	0,98
Facteur 3	0,85
Facteur 4	-0,85
Facteur 5	0,09

Tableau 7

Conclusion

Bien entendu, la variante de l'A.F.C.M. n'exclue pas la technique utilisée dans SPAD mais la complète.

Si l'on a peu de données manquantes et de modalités à faibles effectifs, il est plus intéressant d'utiliser la technique de SPAD qui diagonalise une matrice de plus petite dimension et donne des résultats stables avec un coût calcul réduit. Par contre, la variante de l'A.F.C.M. trouve un compromis, devant de fortes perturbations du tableau des données, entre coût calcul et stabilité des résultats, qui est très avantageux.

Références bibliographiques

(1) BENALI H. (1985) : Stabilité de l'analyse en composantes principales et de l'analyse des correspondances multiples en présence de certains types de perturbations - Méthodes de dépouillement d'enquêtes.
Thèse de 3ème cycle - Université de Rennes 1.

(2) BENZECRI J.P. (1973) : L'Analyse des Données. Tome 2 : L'Analyse des correspondances - Dunod.

(3) ESCOFIER B. (1981) : Traitement de questionnaires avec non-réponses et analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte.
IRISA - Publication Interne n° 146.

(4) LEBART L. (1975) : Validation des résultats en analyse des données. Rapport CREDOC - DGRST.

(5) LEBART L. - MORINEAU A. (1982) : SPAD = Système portable d'analyse des données, CESIA.

(6) MORINEAU A. (1975) : Quelques générateurs de nombres pseudo-aléatoires. Méth. et Prog. Stat. Bull - Techn. du CESIA n° 1.

PI 286 **La tolérance aux fautes dans un système temps-réel à contraintes strictes**
Maryline Silly – 32 pages ; Février 86.

PI 287 **A new statistical approach for the automatic segmentation of continuous speech signals**
Régine André-Obrecht – 38 pages ; Mars 86.

PI 288 **Synthèse sur les réseaux locaux temps-réel**
Philippe Belmans – 40 pages ; Mars 86.

PI 289 **Calcul distribué d'un extrémum et du routage associé dans un réseau quelconque**
Jean-Michel Hélary, Aomar Maddi, Michel Raynal – 36 pages ; Mars 86.

PI 290 **A new matrix multiplication systolic array**
Patrice Quinton, Brigitte Joinnault, Pierrick Gachet – 12 pages ; Avril 86.

PI 291 **Une introduction à quelques techniques du contrôle distribué à travers un exemple**
Noël Plouzeau, Michel Raynal, Jean-Pierre Verjus – 22 pages ; Avril 86.

PI 292 **Distributed Synchronization of Parallel Programs : why and how ?**
Patrice Quinton, Jean-Pierre Verjus – 16 pages ; Avril 86.

PI 293 **Sur l'utilisation des séquences multi-images en robotique**
Lionel Marcé – 16 pages ; Avril 86.

PI 294 **Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs**
Habib Bénali – 16 pages ; Avril 86.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique